Dense Cross-Modal Correspondence Estimation with the Deep Self-Correlation Descriptor

Seungryong Kim, *Member, IEEE,* Dongbo Min, *Senior Member, IEEE,* Stephen Lin, *Member, IEEE,* and Kwanghoon Sohn, *Senior Member, IEEE*

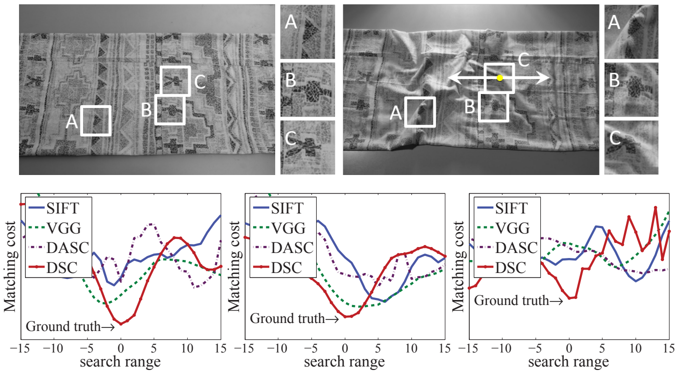
Abstract—We present the deep self-correlation (DSC) descriptor for establishing dense correspondences between images taken under different imaging modalities, such as different spectral ranges or lighting conditions. In this descriptor, local self-similar structure is modeled in a hierarchical manner that yields more precise localization ability and greater robustness to non-rigid image deformations than state-of-the-art descriptors. Specifically, DSC first computes multiple self-correlation surfaces over a local support window for randomly sampled patches, and then builds hierarchical self-correlation surfaces through average pooling. The feature responses on the self-correlation surfaces are then encoded through spatial pyramid pooling in a circular configuration. To better handle geometric variations such as scale and rotation, we further propose the geometry-invariant DSC (GI-DSC) that leverages a multi-scale self-correlation surface and a canonical orientation estimation technique. In contrast to descriptors based on deep convolutional neural networks (CNNs), DSC and GI-DSC are training-free, i.e., handcrafted descriptors, are robust to cross-modal imaging, and cannot be overfitted to the appearance variations of specific modalities. The state-of-the-art performance of DSC and GI-DSC on challenging cases of cross-modal image pairs with photometric and geometric variations is demonstrated through extensive experiments.

Index Terms—Cross-modal correspondence, hierarchical structure, self-correlation, local self-similarity, non-rigid deformation

1 Introduction

Recently in many computer vision and computational photography applications, images captured under different imaging modalities have been used to supplement the data provided in color images. Typical examples of other imaging modalities include infrared [1], [2], [3] and dark flash [4] photography. More broadly, photos taken under different imaging conditions, such as different exposure settings [5], blur levels [6], [7], and illumination [8], can also be considered as cross-modal [9], [10].

Establishing dense correspondences between cross-modal image pairs is essential for combining their disparate information. However, basic visual properties, including color, gradients, and structural similarity, are frequently not shared across cross-modal images, and this degrades matching by conventional feature descriptors [11], [12]. Moreover, geometric variations frequently appear for cross-modal images that are taken under different viewpoints or contain moving objects. Although powerful global optimizers may help to improve the accuracy of correspondence estimation to some extent [13], [14], inherent limitations exist without the use of suitable matching descriptors [15]. The most popular local descriptor is scale invariant feature transform (SIFT) [11], which provides relatively good matching performance when there are small photometric variations. However, conventional descriptors such as SIFT often fail to capture reliable matching evidence in cross-modal images due to their different visual properties [9], [10].



(a) Cost profile of A (b) Cost profile of B (c) Cost profile of C

Fig. 1. Examples of matching cost profiles, computed with different descriptors such as SIFT [11], VGG [16], and DASC [10] along the scan lines of A, B, and C for image pairs under non-rigid deformations and illumination changes. In comparison to other handcrafted and deep CNN-based descriptors, DSC yields more reliable global minima.

Features learned from convolutional neural networks (CNNs) [17], [18], [19], [20], [21], [22], [23], [24] have recently emerged as a robust alternative. However, CNN-based descriptors cannot satisfactorily deal with severe differences in cross-modal appearance, since the shared convolutional kernels across images lead to inconsistent responses [21], [25], similar to conventional handcrafted descriptors. Furthermore, some of these methods are designed for estimating sparse correspondences [22], [23], [26] and cannot in practice provide dense descriptors due to their high computational complexity. Of particular importance, there lacks

- *S. Kim and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea.*
  E-mail: {srkim89, khsohn}@yonsei.ac.kr
- *D. Min is with the Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea.*
  E-mail: dbmin@ewha.ac.kr
- *S. Lin is with Microsoft Research, Beijing 100080, China.*
  E-mail: stevelin@microsoft.com

* *Corresponding author*

Manuscript received Jan. 28, 2019

a benchmark with dense ground-truth correspondences on cross-modal images, making supervised learning of CNNs less feasible for this task. Furthermore, networks trained on small-scale datasets with different modalities may often be overfitted to the appearance variations of specific modalities.

To address the problem of cross-modal appearance and shape changes, feature descriptors have been proposed based on local self-similarity (LSS) [27], which is motivated by the notion that the geometric layout of local internal self-similarities is relatively insensitive to imaging properties. The state-of-the-art descriptor for cross-modal dense correspondence, called dense adaptive self-correlation (DASC) [10], [28], makes use of LSS and has demonstrated high accuracy and speed on cross-modal image pairs. However, DASC suffers from two significant shortcomings. One is its limited discriminative power due to a limited set of patch sampling patterns used for modeling internal self-similarities. In fact, the matching performance of DASC may fall short of CNN-based descriptors on images that share the same modality. The other major shortcoming is that the DASC descriptor does not provide the flexibility to deal with non-rigid deformations, which leads to lower robustness in matching. More recently, a fully convolutional self-similarity (FCSS) descriptor [24], [29] was proposed to formulate LSS within a deep network. However, its application to cross-modal correspondence has not been studied, and it is also vulnerable to non-rigid geometric variations.

In this paper, we introduce a novel descriptor, called deep self-correlation (DSC), that overcomes the shortcomings of DASC [10], [28] and FCSS [24], [29] while providing robust dense cross-modal correspondence. This work is motivated by the observation that local self-similarity can be formulated in a hierarchical structure to enhance localization ability and gain robustness to non-rigid photometric and geometric deformations. Unlike DASC [10], [28] and FCSS [24], [29] that select patch pairs within a support window and calculate the self-similarity between them, DSC computes self-correlation surfaces that more comprehensively encode the intrinsic structure by calculating the self-similarity between randomly selected patches and all of the patches within the support window. These self-correlational responses are aggregated through spatial pyramid pooling in a circular configuration, which yields a representation less sensitive to non-rigid image deformations than the fixed patch selection strategy used in DASC [10], [28] and FCSS [24], [29]. To further enhance localization ability and robustness, we build hierarchical self-correlation surfaces, together with nonlinear and normalization layers. For efficient computation of DSC over densely sampled pixels, we calculate the self-correlation surfaces through fast edge-aware filtering.

Furthermore, to better address geometric variations that may exist across cross-modal image pairs, we propose the geometry-invariant DSC descriptor, called GI-DSC. In formulating this extension, we leverage the observation that geometric deformation fields across cross-modal images can be well approximated locally by a similarity transformation (i.e., translation, rotation, and uniform scale transformation). Specifically, to deal with scale deformations at each pixel, multi-scale self-correlation surfaces are first measured on the image pyramid, and then fused by max-pooling to encode maximal self-similarities of each sampling pattern across scales. Canonical orientations on each pixel are also estimated with the maximum orientation bin weighted by self-correlation surfaces, which are used to build a geometry-invariant descriptor.

Compared to existing CNN-based descriptors [17], [18], [19], [20], [21], [22], [23], [24] (as well as FCSS [24], [29]), DSC requires no training data, since the convolutional kernels are defined using the local self-similarity between pairs of image patches. Fig. 1 illustrates the robustness of DSC for image pairs against non-rigid deformations and illumination changes in comparison to existing handcrafted and even deep CNN-based methods.

In the experimental results, we show that DSC outperforms existing feature descriptors and similarity measures on various benchmarks having photometric and geometric variations: (1) the Middlebury stereo benchmark [30] containing illumination and exposure variations; (2) a cross-modal and cross-spectral dataset [9] including RGB and near-infrared (NIR) images [1], [9], different exposures [5], [9], flash-noflash images [8], blurry images [6], [7], and RGB-depth images [9]; (3) the DaLI benchmark [31] containing non-rigid deformations; (4) the tri-modal human body segmentation benchmark [32] including RGB, depth, and far-infrared (FIR) images; and (5) the DIML benchmark [28] including RGB images with both photometric and geometric variations.

This manuscript extends the conference version of this work [33] through (1) a geometry-invariant extension of DSC, called GI-DSC; (2) an in-depth analysis of DSC and GI-DSC; and (3) an extensive comparative study with state-of-the-art CNN-based descriptors using various datasets. The source code is available online at our project webpage: `http://diml.yonsei.ac.kr/∼srkim/DSC/`.

2 Related Work

2.1 Handcrafted and Learned Feature Descriptors

Conventional gradient-based descriptors, such as SIFT [11] and DAISY [12], as well as intensity comparison-based binary descriptors, such as BRIEF [34], have shown a limited performance in dense correspondence estimation between cross-modal image pairs. Besides these handcrafted features, several attempts have been made using machine learning algorithms to derive features from large-scale datasets [17], [35]. A few of these methods use deep CNNs [36], which have revolutionized image-level classification, to learn discriminative descriptors for local patches. For designing explicit feature descriptors based on a CNN architecture, immediate activations are extracted as the descriptor [17], [18], [19], [20], [21], [22], [23], [24], and have been shown to be effective for this patch-level task. However, even though CNN-based descriptors encode a discriminative structure with a deep architecture, they have inherent limitations in cross-modal image correspondence because they are derived from convolutional layers using shared kernels or volumes [21], [25]. Furthermore, the dearth of ground-truth data for dense cross-modal correspondence presents an obstacle for supervised learning of CNNs in this context.

To estimate cross-modal image correspondences, variants of the SIFT descriptor have been developed [37], but these gradient-based descriptors maintain an inherent limitation similar to SIFT in dealing with image gradients that vary differently between modalities. For illumination invariant correspondences, Wang et al. proposed the local intensity order pattern (LIOP) descriptor [38], but radiometric variations often alter the relative order of pixel intensities. Simo-Serra et al. proposed the deformation and light invariant (DaLI) descriptor [31] to provide high resilience to non-rigid image transformations and illumination changes, but it in practice cannot provide dense descriptors in the image domain due to its heavy computational load. Recently, the cross-spectral similarity model [39], [40] through CNNs has shown improved performance on RGB-NIR correspondence, but it requires supervised learning, thus limiting its applicability to various cross-modal correspondence tasks.

Schechtman and Irani introduced the local self-similarity (LSS) descriptor [27] for the purpose of template matching, and achieved impressive results in object detection and retrieval. By employing LSS, many approaches have tried to solve for cross-modal correspondence [41], [42], [43]. However, none of these approaches scale well to dense matching due to low discriminative power and high complexity. Inspired by LSS, Kim et al. proposed the DASC descriptor to estimate cross-modal dense correspondences [10]. Though it can provide satisfactory performance, it is not able to handle non-rigid deformations and has limited discriminative power due to its fixed patch pooling scheme. More recently, the FCSS descriptor [24] formulated LSS within a fully convolutional network where patch sampling patterns and the self-similarity measure are both learned. Although FCSS improved performance dramatically in semantic correspondence estimation, it is tailored to object-level correspondence estimation, instead of cross-modal image pairs at a scene-level. Moreover, it cannot deal with severe geometric variations which frequently appear across cross-modal images.

2.2 Area-Based Similarity Measures

A popular measure for registration of cross-modal medical images is mutual information (MI) [44], based on the entropy of the joint probability distribution function, but it provides reliable performance only for variations undergoing a global transformation. In [45], this issue is alleviated to some extent by leveraging a locally adaptive weight obtained from SIFT matching but its performance is still limited on cross-modal variation [46]. Although cross-correlation based methods such as adaptive normalized cross-correlation (ANCC) [47] produce satisfactory results for locally linear variations, they are less effective against more substantial modality variations. Irani et al. employed cross-correlation on a Laplacian energy map for measuring multi-sensor image similarity [48], but this exhibits limited performance in general image matching tasks. Robust selective normalized cross-correlation (RSNCC) [9] was proposed for dense alignment between cross-modal images, but as an intensity based measure it can still be sensitive to cross-modal variations. DeepMatching [49] was proposed to compute dense correspondences by employing a hierarchical pooling scheme like in a CNN, but it is not designed to handle cross-modal matching.

2.3 Geometry-Invariant Correspondence Estimation

To alleviate geometric variation problems in establishing dense correspondences, many methods have been proposed based on SIFT flow (SF) [13] optimization, including deformable spatial pyramid (DSP) [14], scale-less SIFT flow (SLS) [50], scale-space SIFT flow (SSF) [51], and generalized DSP (GDSP) [52]. However, the large search spaces for establishing geometry-invariant dense correspondence make computational complexity a critical limitation of these methods. Generalized PatchMatch (GPM) [53] was proposed for efficient matching based on a randomized search scheme. DAISY Filter Flow (DFF) [54], which utilizes the DAISY descriptor [12] with the PatchMatch Filter (PMF) [55], was proposed to provide geometric invariance. However, their weak spatial smoothness often induces mismatched results. While the aforementioned methods have attempted to address the problem from an optimization perspective, various geometry-invariant descriptors have also been developed for geometry-invariant correspondence estimation. The scale invariant descriptor (SID) [56] was proposed to encode geometric robustness in the descriptor itself, but it does not deal with multi-modal matching. A segmentation-aware approach [57] was presented to provide geometric robustness for descriptors, e.g., SIFT [11] or SID [56], but it can have a negative effect on the discriminative power of the descriptor. More recently, as an extension of DASC, geometry-invariant DASC (GI-DASC) [28] employed DASC in a superpixel-based representation with estimated geometric fields. Although it provides improved robustness to geometric variations, it inherits the limitations of DASC, and its performance is sensitive to superpixel segmentation accuracy.

3 Background

Let us define an image as $f_i : \mathcal{I} \to \mathbb{R}$ for pixel i, where $\mathcal{I} \subset \mathbb{N}^2$ is a discrete image domain. Given the image f_i, a dense descriptor $\mathcal{D}_i : \mathcal{I} \to \mathbb{R}^L$ with a feature dimension of L is defined on a local support window $\mathcal{R}_i$.

Unlike conventional descriptors which rely on basic visual properties such as color and gradients [11], [12], LSS-based descriptors provide robustness to different imaging modalities since internal local self-similarities are preserved across cross-modal image pairs [10], [24], [27]. As shown in Fig. 2(a), LSS [27] first computes the self-correlation surface, discretizes the correlation surface into log-polar bins, and then stores the maximum correlation value of each bin. Formally, it generates an $L^{\text{lss}} \times 1$ feature vector $\mathcal{D}_i^{\text{lss}} = \bigcup_l d_i^{\text{lss}}(l)$ for $l \in \{1, ..., L^{\text{lss}}\}$, with $d_i^{\text{lss}}(l)$ computed as

$$d_i^{\text{lss}}(l) = \max_{j \in \mathcal{B}_i(l)} \{\exp(-\mathcal{S}(i,j)/\sigma_c)\}, \tag{1}$$

where a log-polar bin is defined as $\mathcal{B}_i = \{j | j \in \mathcal{R}_i, \rho_{r-1} < |i-j| \leq \rho_r, \phi_{a-1} < \angle(i-j) \leq \phi_a\}$ with a log radius ρ_r for $r \in \{1, \cdots, N_\rho\}$ and a quantized angle ϕ_a for $a \in \{1, \cdots, N_\phi\}$ with $\rho_0 = 0$ and $\phi_0 = 0$. Each pair of r and a is associated with a unique index l. $\mathcal{S}(i,j)$ is the correlation between patches $\mathcal{F}_i$ and $\mathcal{F}_j$, computed using the sum of squared differences (SSD) [27]. Though LSS provides

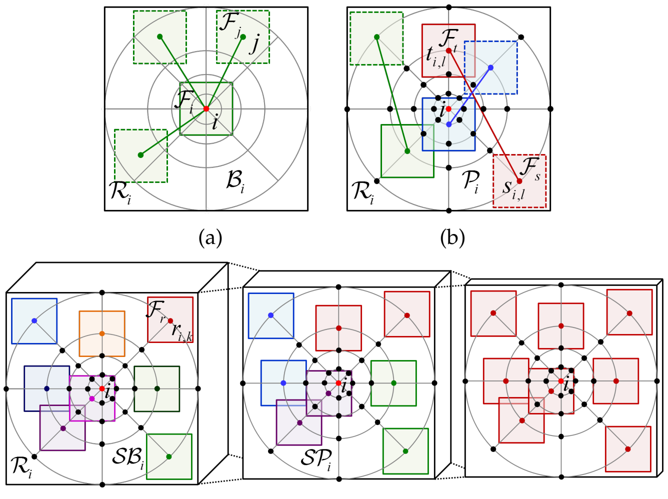


Fig. 2. Illustration of (a) LSS [27] using center-biased dense max pooling, (b) DASC [10] and FCSS [24] using patch-wise receptive field pooling, and (c) our DSC. Boxes, formed by solid and dotted lines, depict source and target patches. DSC incorporates circular spatial pyramid pooling on hierarchical self-correlation surfaces.

robustness to modality variations, matching details are not well preserved and its significant computation does not scale well for estimating dense correspondences.

Inspired by LSS [27], DASC [10] encodes the self-similarity between patch-wise receptive fields sampled from a log-polar circular point set $\mathcal{P}_i$ as shown in Fig. 2(b). It is defined such that $\mathcal{P}_i = \{j | j \in \mathcal{R}_i, |i-j| = \rho_r, \angle(i-j) = \phi_a\}$, which has a higher density of points near the center pixel, similar to DAISY [12]. DASC is encoded with a set of similarities between patch pairs of sampling patterns selected from $\mathcal{P}_i$ such that $\mathcal{D}_i^{\text{dasc}} = \bigcup_l d_i^{\text{dasc}}(l)$ for $l \in \{1, ..., L^{\text{dasc}}\}$:

$$d_i^{\text{dasc}}(l) = \exp(-(1 - |\mathcal{C}(s_{i,l}, t_{i,l})|)/\sigma_c), \tag{2}$$

where $s_{i,l}$ and $t_{i,l}$ are the l^{th} selected sampling pattern from $\mathcal{P}_i$. Patch-wise similarity is computed with an exponential function with a bandwidth of σ_c, which has been widely used for robust estimation [58]. Here, an absolute value of $\mathcal{C}(s_{i,l}, t_{i,l})$ is used for mitigating the effect of intensity reverses. $\mathcal{C}(s_{i,l}, t_{i,l})$ is computed using an adaptive self-correlation measure inspired by [47]. Although the DASC descriptor has shown satisfactory results for cross-modal dense correspondence estimation, its randomized receptive field pooling has limited descriptive power and does not accommodate non-rigid deformations.

4 The DSC Descriptor

4.1 Motivation and Overview

Inspired by DASC [10], [28], our DSC descriptor also represents an adaptive self-correlation measure between two patches within a local support window. However, we adopt a different strategy where hierarchical self-correlation surfaces are built through the spatial aggregation of self-correlation responses on a single level, and feature responses more comprehensively encode local self-similar structure to improve localization ability and robustness to non-rigid image deformation (Sec. 4.2). Densely sampled descriptors are efficiently computed over an entire image using a method

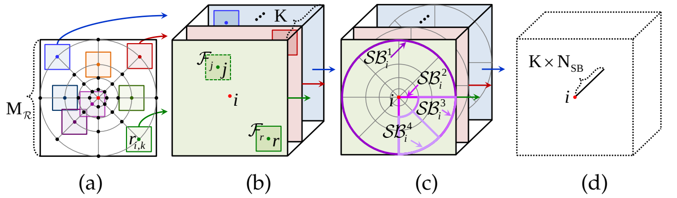


Fig. 3. Computation of the single self-correlation (SSC) descriptor for (a) a local support window with random samples. (b) For each random patch, a self-correlation surface is computed using an adaptive self-correlation measure. (c) A self-correlation response is then obtained through circular spatial pyramid pooling (C-SPP). (d) The responses from C-SPP are concatenated into a feature vector.

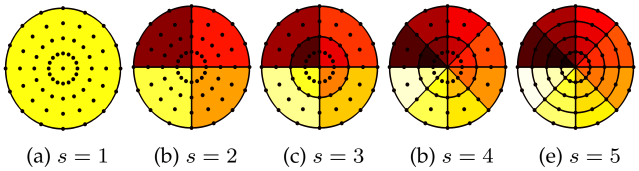


Fig. 4. Examples of circular spatial pyramidal bins $\mathcal{SB}$. The total number of bins is $N_{\mathcal{SB}} = \sum_{s=2}^{S} 2^s + 1$, where S represents the pyramid level.

based on fast edge-aware filtering (Sec. 4.3). We further build hierarchical self-correlation surfaces to enhance the robustness of the descriptor (Sec. 4.4). Finally, to alleviate problems caused by geometric variations, scale and/or rotation, deformations on each pixel are addressed in building the GI-DSC descriptor (Sec. 4.5). Fig. 2(c) illustrates the DSC descriptor, which incorporates circular spatial pyramid pooling on hierarchical self-correlation surfaces.

4.2 SSC: Single Self-Correlation

To overcome the limitations of self-similarity in DASC [10] and FCSS [24] descriptors, our approach builds hierarchical self-correlation surfaces, where feature responses are obtained through circular spatial pyramid pooling. We start by describing a single-layer version of DSC, which we refer to as single self-correlation (SSC).

4.2.1 *Self-Correlations*

To build multiple self-correlation surfaces in SSC, we randomly select K points from a log-polar point set $\mathcal{P}_i$ defined within a local support window. We then convolve a patch $\mathcal{F}_{r_{i,k}}$ centered at the k-th point $r_{i,k}$ with all patches $\mathcal{F}_j$, defined for $k \in \{1, ..., K\}$ and $j \in \mathcal{R}_i$ as shown in Fig. 3(b). Similar to DASC [10], the similarity $\mathcal{C}(r_{i,k}, j)$ between patch pairs is measured using an adaptive self-correlation, which is known to be effective in addressing cross-modality. With (i, k) omitted for simplicity, the $\mathcal{C}(r, j)$ is computed as follows:

$$\mathcal{C}(r,j) = \frac{\sum\limits_{r',j'} \omega_{r,r'}\omega_{j,j'}(f_{r'} - G_r^r)(f_{j'} - G_j^j)}{\sqrt{\sum\limits_{r'} \{\omega_{r,r'}(f_{r'} - G_r^r)\}^2} \sqrt{\sum\limits_{j'} \{\omega_{j,j'}(f_{j'} - G_j^j)\}^2}}, \tag{3}$$

where $r' \in \mathcal{F}_r$ and $j' \in \mathcal{F}_j$, and $G_r^r = \sum_{r'} \omega_{r,r'} f_{r'}$ and $G_j^j = \sum_{j'} \omega_{j,j'} f_{j'}$ represent weighted averages of $f_{r'}$ and $f_{j'}$. Similar to DASC [10], the weight $\omega_{r,r'}$ represents how similar two pixels r and r' are, and is normalized, i.e., $\sum_{r'} \omega_{r,r'} = 1$. It may be defined using any form of edge-aware weighting [59], [60], which increases precision in describing self-similarities and boosts matching performance.

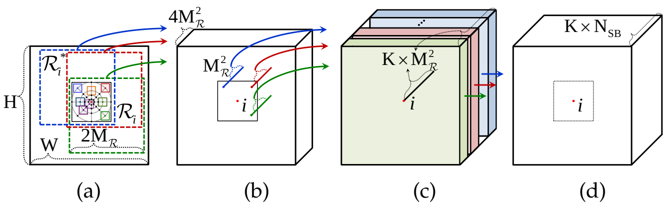


Fig. 5. Efficient computation of self-correlation surfaces in an image: (a) An image f_i with a doubled support window $\mathcal{R}_i^*$ and random samples. (b) A 1-D vector representation of a self-correlation surface. (c) Self-correlation surfaces. (d) Self-correlation responses after C-SPP. With edge-aware filtering and response reformulation, self-correlation responses are computed efficiently in a dense manner.

4.2.2 Circular Spatial Pyramid Pooling

To encode the feature responses on the self-correlation surface, we propose a circular spatial pyramid pooling (C-SPP) scheme, which pools the responses within each hierarchical spatial bin, similar to spatial pyramid pooling (SPP) [25], [61], [62] but in a circular configuration. Note that many existing descriptors also adopt a circular pooling scheme, which brings greater robustness because of its higher pixel density near the central pixel [12], [27], [34]. We further encode more structure information with a C-SPP.

The circular pyramidal bins $\mathcal{SB}_i(u)$ are defined from log-polar circular bins $\mathcal{B}_i$, where u indexes all pyramidal levels $s \in \{1, ..., S\}$ and all bins in each level s as in Fig. 4. The circular pyramidal bin at the top of pyramid, i.e., $s = 1$, encompasses all of the bins $\mathcal{B}_i$. The second level, i.e., $s = 2$, is defined by dividing $\mathcal{B}_i$ into quadrants. For lower pyramid levels, i.e., $s > 2$, the circular pyramidal bins are defined differently according to whether s is odd or even. For an odd s, the bins are defined by dividing bins in the upper level into two parts along the radius. For an even s, they are defined by dividing bins in the upper level into two parts with respect to the angle. The set of all circular pyramidal bins is denoted as $\mathcal{SB}_i = \bigcup_u \mathcal{SB}_i(u)$ for $u \in \{1, ..., N_{\mathcal{SB}}\}$, where the number of circular spatial pyramid bins is defined as $N_{\mathcal{SB}} = \sum_{s=2}^{S} 2^s + 1$.

As illustrated in Fig. 3(c), the feature responses are finally max-pooled on the circular pyramidal bins $\mathcal{SB}_i(u)$ of each self-correlation surface $\mathcal{C}(r_{i,k}, j)$, yielding the following feature response:

$$g_{i,k}(u) = \max_{j \in \mathcal{SB}_i(u)} \{\mathcal{C}(r_{i,k}, j)\}. \tag{4}$$

This max-pooling is repeated for all k and u, yielding accumulated correlation responses $g_i^{\text{ssc}}(l) = \bigcup_{\{k,u\}} g_{i,k}(u)$ where l indexes over all k and u.

Interestingly, LSS [27] also uses a max pooling strategy to mitigate the effects of non-rigid image deformation. However, the max pooling in the single-scale self-correlation surface of LSS [27] loses fine-scale matching details as reported in [10]. By contrast, our descriptor employs circular spatial pyramid pooling in a multi-scale self-correlation surface that provides a more discriminative representation of self-similarities, thus maintaining fine-scale matching details as well as providing robustness to non-rigid deformations.

4.2.3 Non-linear Mapping and Normalization

The feature responses are passed through a non-linear mapping and a normalization to mitigate the effects of outliers.

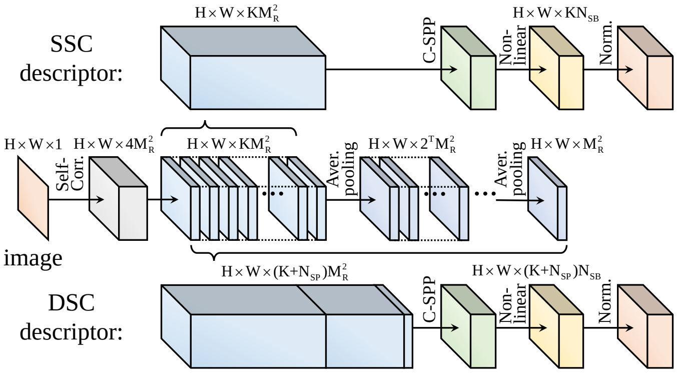


Fig. 6. Visualization of the SSC and DSC descriptors. Our architecture consists of a hierarchical self-correlational layer, circular spatial pyramid pooling layer, non-linear gating layer, and normalization layer.

With accumulated correlation responses $g_i^{\text{ssc}}(l)$, the SSC descriptor $\mathcal{D}_i^{\text{ssc}} = \bigcup_l d_i^{\text{ssc}}(l)$ is computed for $l \in \{1, ..., L^{\text{ssc}}\}$ through a non-linear mapping:

$$d_i^{\text{ssc}}(l) = \exp(-(1 - |g_i^{\text{ssc}}(l)|)/\sigma_c), \tag{5}$$

where σ_c is the Gaussian kernel bandwidth. The features obtained from the SSC descriptor are of size $L^{\text{ssc}} = K \times N_{\mathcal{SB}}$. Finally, $d_i^{\text{ssc}}(l)$ for each pixel i is normalized with an L-2 norm for all l.

4.3 Efficient Computation for Dense Description

The most time-consuming part of SSC is in constructing self-correlation surfaces $\mathcal{C}(r_{i,k}, j)$ for k and j, where $K \times M_{\mathcal{R}}^2$ computations of (3) are needed for each pixel i. Straightforward computation of a weighted summation using ω in (3) would require considerable processing with a computational complexity of $O(IM_{\mathcal{F}}KM_{\mathcal{R}}^2)$, where $I = H \times W$ represents the size of an image (height H and width W). To expedite processing, we pre-compute the self-correlation surfaces within a larger local support window, accelerated by utilizing fast edge-aware filtering [59], [60].

First of all, we compute $\mathcal{C}(r_{i,k}, j)$ efficiently by rearranging all sampling patterns $(r_{i,k}, j)$ into reference-biased pairs $(i, h) = (i, i + r_{i,k} - j)$. Similar to DASC [10], $\mathcal{C}(i, h)$ can be expressed in an approximate form[1] as

$$\hat{\mathcal{C}}(i, h) = \frac{\sum_{i',h'} \omega_{i,i'}(f_{i'} - G_i^i)(f_{h'} - G_h^i)}{\sqrt{\sum_{i'} \omega_{i,i'}(f_{i'} - G_i^i)^2} \sqrt{\sum_{i',h'} \omega_{i,i'}(f_{h'} - G_h^i)^2}}, \tag{6}$$

where $G_i^i = \sum_{i'} \omega_{i,i'} f_{i'}$ and $G_h^i = \sum_{i',h'} \omega_{i,i'} f_{h'}$. For faster computation, it can be expressed as follows [10]:

$$\hat{\mathcal{C}}(i, h) = \frac{\mathcal{G}_{ih}^i - G_i^i \cdot G_h^i}{\sqrt{G_{i^2}^i - (G_i^i)^2} \cdot \sqrt{G_{h^2}^i - (G_h^i)^2}}, \tag{7}$$

where $G_{ih}^i = \sum_{i',h'} \omega_{i,i'} f_{i'} f_{h'}$, $G_{i^2}^i = \sum_{i'} \omega_{i,i'} f_{i'}^2$, and $G_{h^2}^i = \sum_{i',h'} \omega_{i,i'} f_{h'}^2$. $\hat{\mathcal{C}}(i, h)$ can be efficiently computed using any form of fast edge-aware filter [59], [60] with a complexity of $O(IKM_{\mathcal{R}}^2)$. $\mathcal{C}(r, j)$ is then simply obtained from $\hat{\mathcal{C}}(i, h)$ by re-indexing sampling patterns [10].

[1] As shown in [10], there exists a marginal performance difference between the asymmetric self-correlation measure in (7) and original one in (3).

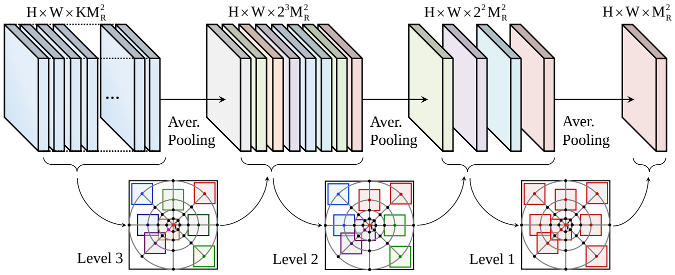


Fig. 7. Visualization of average pooling in DSC. The hierarchical self-correlation surfaces are sequentially aggregated using average pooling from the bottom to the top of the circular pyramidal point set.

Though we remove the computational dependency on patch size $M_{\mathcal{F}}$, $K \times M_{\mathcal{R}}^2$ computations of (7) are still needed to obtain the self-correlation surfaces, where many sampling pair computations for i and h are repeated with respect to $i-h$. To avoid such redundancy, we first compute a self-correlation surface $\mathcal{C}(i,h)$ for $h \in \mathcal{R}_i^*$ with a doubled local support window $\mathcal{R}_i^*$ of size $2 \times M_{\mathcal{R}}$. A doubled local support window is used because (7) is computed with patch $\mathcal{F}_h$, and the minimum support window size for $\mathcal{R}_i^*$ to cover all samples within $\mathcal{R}_i$ is $2 \times M_{\mathcal{R}}$ as shown in Fig. 5(b). After the self-correlation surface for $\mathcal{R}_i^*$ is computed once over the image domain, $\mathcal{C}(r,j)$ can be extracted through an index mapping process, where the indexes for $\mathcal{R}_{i-r_{i,k}}$ are estimated from $\mathcal{R}_i^*$. With this strategy, the computational complexity of constructing self-correlation surfaces becomes $O(I4M_{\mathcal{R}}^2)$, which is smaller than $O(IKM_{\mathcal{R}}^2)$ as $4 \ll K$.

4.4 DSC: Deep Self-Correlation

So far, we have discussed how to build the self-correlation surface on a single level. In this section, we extend this idea by encoding self-similar structures at multiple levels. DSC is defined similarly to SSC, except that average pooling is executed before C-SPP (see Fig. 6). With self-correlation surfaces, we perform the average pooling on circular pyramidal point sets. In comparison to the self-correlations just from a single patch, the spatial aggregation of self-correlation responses is clearly more robust, and it requires only marginal computational overhead over SSC. The strength of such a hierarchical aggregation has also been shown in [49].

Specifically, to build the hierarchical self-correlation surface through average pooling, we first define the circular pyramidal point sets $\mathcal{SP}_i(v)$ from log-polar circular point sets $\mathcal{P}_i$, where v indexes all pyramidal levels $t \in \{1,...,T\}$ and all points in each level t. In the average pooling, the circular pyramidal bins used in C-SPP are re-used such that $\mathcal{SP}_i(v) = \{j | j \in \mathcal{P}_i, j \in \mathcal{SB}_i(u)\}$, thus $T = S$. As shown in Fig. 7, deep self-correlation surfaces are defined by aggregating $\mathcal{C}(r_{i,k},j)$ for all $r_{i,k}$ patches determined on each $\mathcal{SP}_i(v)$ such that

$$\mathcal{C}(v,j) = \sum_{r_{i,k} \in \mathcal{SP}_i(v)} \mathcal{C}(r_{i,k},j)/N_v, \tag{8}$$

which is defined for all v, and N_v is the number of $r_{i,k}$ patches within $\mathcal{SP}_i(v)$. The hierarchical self-correlation surfaces are sequentially aggregated using average pooling from the bottom to the top of the circular pyramidal point set. After computing hierarchical self-correlational aggregations, DSC employs C-SPP as well as non-linear and normalization layers, similar to SSC as presented in Sec. 4.2. A hierarchical self-correlation response $g_{i,v}(u)$ is computed using C-SPP as

$$g_{i,v}(u) = \max_{j \in \mathcal{SB}_i(u)} \{\mathcal{C}(v,j)\}. \tag{9}$$

An accumulated self-correlation response is then built from $g_{i,k}(u)$ in (4) and from $g_{i,v}(u)$ in (9) such that $g_i^{\mathrm{dsc}}(l) = \bigcup_{\{k,v,u\}} \{g_{i,k}(u), g_{i,v}(u)\}$ where l indexes over all k, v, and u. Our DSC descriptor $\mathcal{D}_i^{\mathrm{dsc}} = \bigcup_l d_i^{\mathrm{dsc}}(l)$ is then built from $g_i^{\mathrm{dsc}}(l)$ through a non-linear gating layer as in (5) for $l \in \{1,...,L^{\mathrm{dsc}}\}$ with $L^{\mathrm{dsc}} = (K + N_{\mathcal{SP}})N_{\mathcal{SB}}$. Finally, $d_i^{\mathrm{dsc}}(l)$ for each pixel i is normalized with an L-2 norm for all l.

Algorithm 1: Deep Self-Correlation (DSC) Descriptor

```
Input: image $f_i$, random samples $r_{i,k}$
Output: DSC descriptor $\mathcal{D}_i^{\mathrm{dsc}}$
Parameters: number of circular pyramidal bins (points) $N_{\mathcal{SB}}(N_{\mathcal{SP}})$
Compute $\hat{\mathcal{C}}(i,h)$ for a doubled support window $\mathcal{R}_i^*$ by using (7).
Compute $\mathcal{C}(r_{i,k},j)$ from $\hat{\mathcal{C}}(i,h)$ according to the index mapping.
for $v = 1 : N_{\mathcal{SP}}$ do
    /* Hierarchical Aggregation using Average Pooling */
    Determine a circular pyramidal point $\mathcal{SP}_i(v)$.
    Compute $\mathcal{C}(v,j)$ by using average pooling
    for $\mathcal{SP}_i(v)$ on $\mathcal{C}(r_{i,k},j)$.
end for
for $u = 1 : N_{\mathcal{SB}}$ do
    /* Hierarchical Spatial Aggregation using C-SPP */
    Determine a circular pyramidal bin $\mathcal{SB}_i(u)$.
    Compute $g_{i,k}(u)$ and $g_{i,v}(u)$ by using C-SPP
    on each $\mathcal{SB}_i(u)$ from $\mathcal{C}(r_{i,k},j)$ and $\mathcal{C}(v,j)$, respectively.
end for
Build hierarchical self-correlation responses $g_i^{\mathrm{dsc}}(l)$ from
$g_{i,k}(u)$ and $g_{i,v}(u)$.
Compute a DSC descriptor $\mathcal{D}_i^{\mathrm{dsc}} = \bigcup_l d_i^{\mathrm{dsc}}(l)$,
followed by L-2 normalization.
```

4.5 Geometry-invariant DSC

It is known that LSS-based descriptors [10], [24], [27], [63] provide geometric invariance to some extent thanks to the log-polar binning of a self-correlation surface. However, under more significant geometric variations, existing LSS-based descriptors including DSC do not provide satisfactory performance without an explicit module to consider geometric variations. To overcome this issue, we propose the geometry-invariant DSC (GI-DSC) that explicitly addresses scale and/or rotation deformations between cross-modal images. The underlying assumption is that geometric deformation fields across cross-modal images can be locally well approximated by a similarity transformation (i.e., translation, rotation, and uniform scale transformation).

4.5.1 Scale-Invariant Self-Correlations

An explicit scale estimation technique using a scale-space of conventional feature detectors such as SIFT [11] is sensitive to cross-modal deformation as exemplified in [28]. We observe that the maximal self-correlation of each sampling pattern across multiple scales remains consistent with respect to image scale.

Specifically, to compute a multi-scale self-correlation surface, we first build the Gaussian image pyramid $f_i^m = f_i * \varrho_m$ for $m = \{1,...,M\}$, where ϱ_m is the m-th Gaussian kernel and M is the number of pyramid levels. For each

Algorithm 2: Geometry-Invariant DSC (GI-DSC) Descriptor

```
Input: image $f_i$, random samples $r_{i,k}$
Output: GI-DSC descriptor $\mathcal{D}_i^{\mathrm{gi-dsc}}$
Parameters: number of circular pyramidal bins (points) $N_{\mathcal{SB}}(N_{\mathcal{SP}})$
    /* Scale-Invariance */
1 : Compute the Gaussian image pyramid $f_i^m = f_i * \varrho_m$.
2 : Compute $\hat{\mathcal{C}}^m(i,h)$ for $f_i^m$ using (10).
3 : Estimate $\hat{\mathcal{C}}^{\mathrm{si}}(i,h)$ using a max-pooling as in (11).
    /* Rotation-Invariance */
4 : Construct $l_{\mathrm{hist}}(i,\theta_a)$ with $\hat{\mathcal{C}}^{\mathrm{si}}(i,h)$ using (12).
5 : Estimate the orientation $\theta_i$ for each pixel $i$ from $l_{\mathrm{hist}}(i,\theta_a)$.
6 : Filter out the orientation $\theta_i$ to provide smooth geometric fields.
7 : Transform $r_{i,k}$, $\mathcal{SB}_i$, and $\mathcal{SP}_i$ according to $\theta_i$.
8 : Through Step 2-8 in Algorithm 1, compute a GI-DSC descriptor
    such that $\mathcal{D}_i^{\mathrm{gi-dsc}} = \bigcup_l d_i^{\mathrm{gi-dsc}}(l)$.
```

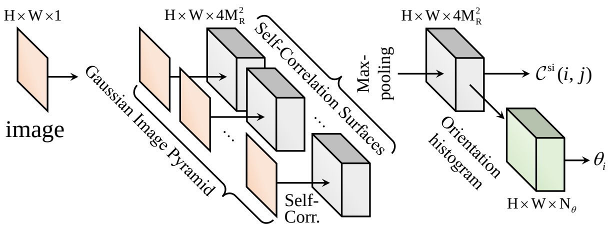


Fig. 8. Visualization of geometry-invariance in GI-DSC. To provide scale invariance, our approach measures multi-scale self-correlation surfaces, and fuses them by max-pooling. Moreover, canonical orientation fields for each pixel are estimated to provide orientation invariance.

image pyramid level f_i^m, we measure the asymmetric self-correlation $\hat{\mathcal{C}}^m(i,h)$, similar to (7), such that

$$\hat{\mathcal{C}}^m(i,h) = \frac{G_{ih}^{i,m} - G_i^{i,m} \cdot G_h^{i,m}}{\sqrt{G_{i^2}^{i,m} - (G_i^{i,m})^2} \cdot \sqrt{G_{h^2}^{i,m} - (G_h^{i,m})^2}}, \tag{10}$$

where $G_{ih}^{i,m}$, $G_i^{i,m}$, $G_h^{i,m}$, $G_{i^2}^{i,m}$, and $G_{h^2}^{i,m}$ are measured for each image pyramid level f_i^m. The scale-invariant self-correlation is then computed by max-pooling as follows:

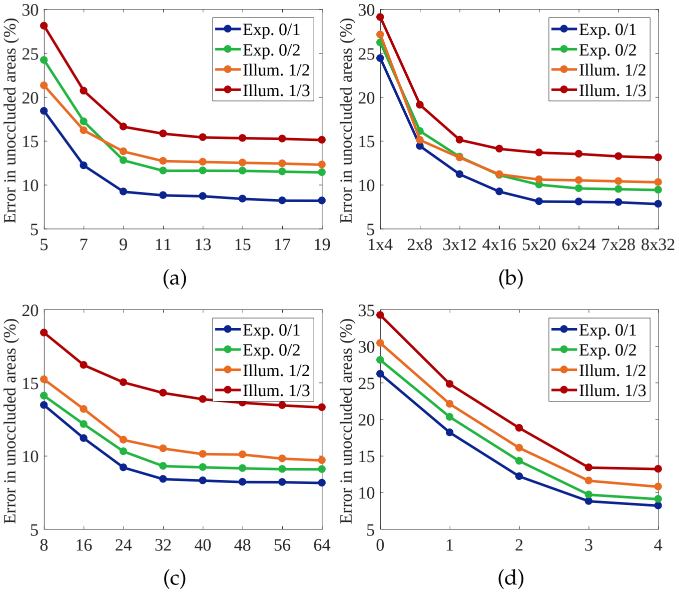
$$\hat{\mathcal{C}}^{\mathrm{si}}(i,h) = \max_{m \in \{1,...,M\}} \{\hat{\mathcal{C}}^m(i,h)\}. \tag{11}$$

4.5.2 *Orientation Estimation for Rotation Invariance*

Similar to scale invariance, rotation invariance also can be achieved by using images under multiple orientations. However, such a technique would dramatically increase computational complexity, a function of the product between the number of scales and rotations. Furthermore, our initial experiments indicated that the localization ability of the descriptor around object boundaries is degraded substantially. Fortunately, unlike scale, the orientation field on each pixel can be easily determined from the maximum among orientation bins weighted by (pre-computed) self-correlations. By transforming the randomly sampled points, the circular pyramidal bins, and the circular pyramidal points according to the estimated orientation, our descriptor can provide rotation invariance on each pixel with only marginal computational overhead.

Specifically, an orientation θ_i of each pixel i is found by constructing a histogram with angles $\angle(i-h)$ for $h \in \mathcal{R}_i^*$ weighted with self-correlations $\hat{\mathcal{C}}(i,h)$ such that

$$l_{\mathrm{hist}}(i,\theta_a) = \sum_{h \in \mathcal{H}_i(a)} \hat{\mathcal{C}}(i,h)/N_a, \tag{12}$$



(a) (b) (c) (d)

Fig. 9. Component analysis of DSC on the Middlebury benchmark [30] for varying parameter values, such as (a) support window size $M_{\mathcal{R}}$, (b) number of log-polar circular point $N_\rho \times N_\phi$, (c) number of random samples K, and (d) level of circular spatial pyramid S. In each experiment, all other parameters are fixed to the initial values.

where $\mathcal{H}_i(a) = \{h | h \in \mathcal{R}_i^*, \theta_{a-1} < \angle(i-h) \le \theta_a\}$ and a quantized angle θ_a for $a \in \{1,...,N_\theta\}$, and N_a is the number of samples in $\mathcal{H}_i(a)$. We then simply choose the main orientation for each pixel corresponding to the most heavily-weighted bin in the histogram, *i.e.*, $\theta_i = \mathrm{argmax}_{\theta_a} l_{\mathrm{hist}}(i,\theta_a)$. Moreover, based on the observation that the geometric deformation fields tend to vary smoothly except at object boundaries [24], [64], the estimated orientation θ_i for each pixel i is propagated to neighboring pixels using a fast (color-guided) global image filter [65] to correct erroneous rotation fields.

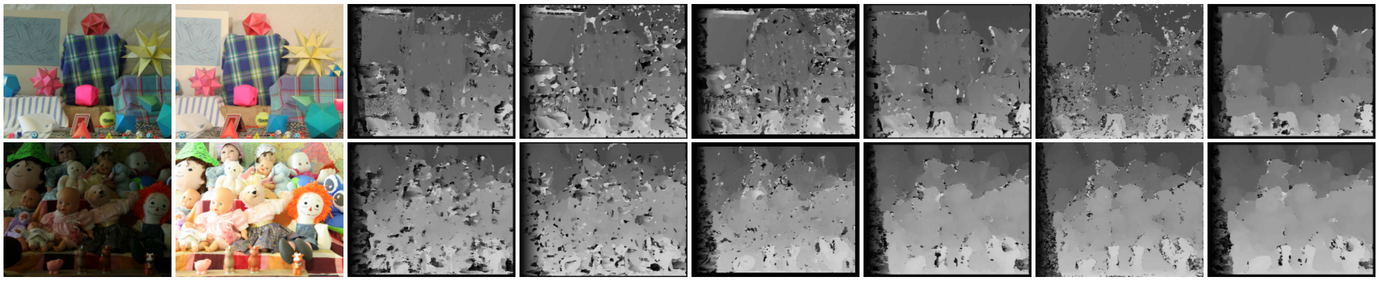
To provide rotation invariance to the DSC descriptor in Sec. 4.4, we transform randomly sampled points $r_{i,k}$, the circular pyramidal bins $\mathcal{SB}_i$ and the circular pyramidal points $\mathcal{SP}_i$ according to estimated rotation θ_i, and then build the DSC descriptor similarly to Fig. 6. By incorporating both scale- and rotation-invariance from Sec. 4.5.1 and Sec. 4.5.2 within the DSC descriptor, we obtain the GI-DSC descriptor with geometric invariance as well as cross-modal robustness. Fig. 8 illustrates this geometry invariance in the GI-DSC descriptor.

5 Experimental Results and Discussion

5.1 Experimental Settings

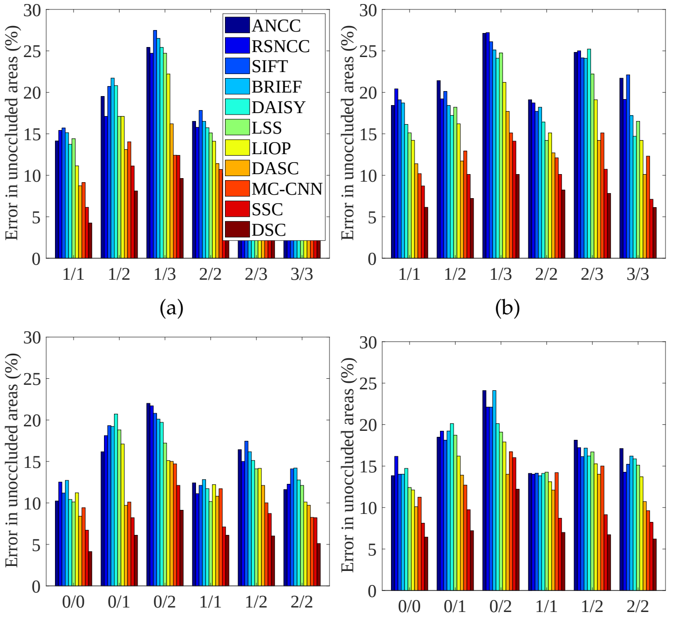
In our experiments, DSC and GI-DSC are implemented with the following fixed parameter settings for all datasets: $\{M_{\mathcal{R}}, M_{\mathcal{F}}, \sigma_c, K, S, T\} = \{9, 5, 0.5, 32, 3, 3\}$, $\{N_\rho, N_\phi\} = \{4, 16\}$, and $\{N_\theta, M\} = \{32, 4\}$. The dimension of SSC and DSC (or GI-DSC) are fixed to 416 and 585, respectively. We choose the guided filter (GF) for edge-aware filtering in (7), with a smoothness parameter of $\epsilon = 0.03^2$. We implement the DSC and GI-DSC descriptors in Matlab/C++ on an Intel Core i7-3770 CPU at 3.40 GHz.

In the following, our DSC and GI-DSC descriptors are compared to other state-of-the-art handcrafted descriptors (SIFT [11], DAISY [12], BRIEF [34], LIOP [38], DaLI [31],



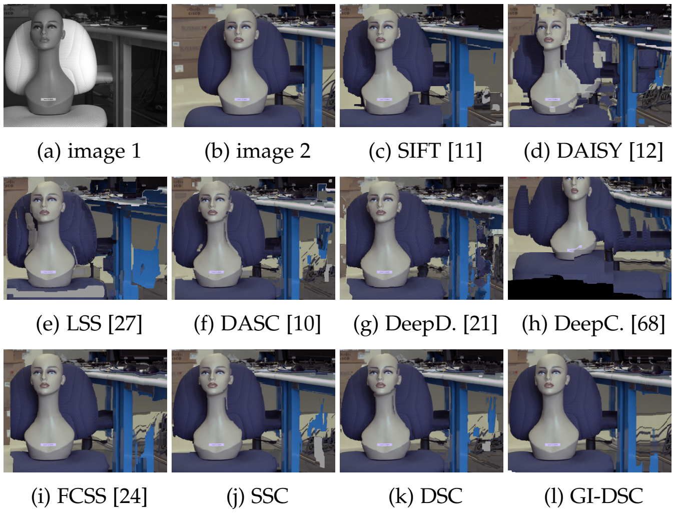
(a) image 1 (b) image 2 (c) ANCC [47] (d) SIFT [11] (e) LSS [27] (f) DASC [10] (g) MCCNN [66] (h) DSC

Fig. 10. Comparison of disparity estimations for *Moebius* and *Dolls* image pairs on the Middlebury benchmark [30] across illumination combination '1/3' and exposure combination '0/2', respectively. Compared to other methods, DSC estimates more accurate and edge-preserved disparity maps.



(a) (b) (c) (d)

Fig. 11. Average bad-pixel error rate on the Middlebury benchmark [30] with illumination and exposure variations. Optimization was done by GC in (a), (b), and by WTA in (c), (d). SSC and DSC descriptors show the best performance with the lowest error rate.



(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]

(e) LSS [27] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [68]

(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 12. Dense correspondence evaluations for RGB-NIR image pairs on cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

LSS [27], SegSIFT [57], SegSID [57], DASC [10], and GI-DASC [28]), recent CNN-based descriptors (MC-CNN [66], VGG[2] [16], FCSS [24], MatchNet (MatchN.) [26], Deep Compare (DeepC.) [68], Deep Descriptor (DeepD.) [21], Learned Invariant Feature Transform (LIFT) [22], L2-Net [23], and Quadruplet Network (Q-Net) [40][3]), and area-based approaches using handcrafted similarity measures (ANCC [47] and RSNCC [9]). Furthermore, to evaluate the performance gain with a hierarchical structure, we compared SSC and DSC. Optimization for all descriptors and similarity measures was done using WTA and SIFT flow (SF) with hierarchical dual-layer belief propagation [13], for which the code is publicly available.

5.2 Ablation Study

The performance of DSC is exhibited in Fig. 9 for varying parameter values, including support window size $M_{\mathcal{R}}$, number of log-polar circular points $N_\rho \times N_\phi$, number of random samples K, and levels of the circular spatial pyramid S. Fig. 9(c) and (d) demonstrate the effectiveness of self-correlation surfaces and hierarchical structures. For a quantitative analysis, we measured the average bad-pixel error rate on the Middlebury benchmark [30]. With a larger support window $M_{\mathcal{R}}$, the matching quality improves rapidly until about 9×9. $N_\rho \times N_\phi$ influences the performance of circular pooling, which is found to plateau at 4×16. Using a larger number of random samples K yields better performance since DSC encodes more information. The levels of circular spatial pyramid S also affect the amount of encoding. Based on these experiments, we set $K = 32$ and $S = 3$ in consideration of efficiency and robustness.

5.3 Middlebury Stereo Benchmark

We first evaluated SSC and DSC on the Middlebury stereo benchmark [30], which contains illumination and exposure variations. In the experiments, the illumination (exposure) combination '1/3' indicates that two images were captured under the 1^{st} and 3^{rd} illumination (exposure) conditions. For quantitative evaluation, we measured the bad-pixel error rate in non-occluded areas of disparity maps [30].

Fig. 10 shows the disparity maps estimated under severe illumination and exposure variations with winner-takes-all (WTA) optimization. Fig. 11 displays the average bad-pixel error rates of disparity maps obtained under illumination or exposure variations, with graph-cut (GC) [69] and WTA

[2] In 'VGG', ImageNet pretrained VGG-Net [16] from the bottom conv1 to the conv3-4 layer were used with L_2 normalization [67].

[3] Since MatchN. [26], DeepC. [68], DeepD. [21], LIFT [22], L2-Net [23], and Q-Net [40] were developed for sparse correspondence, sparse descriptors were first built by forward-propagating images through networks and then upsampled.

TABLE 1
Average error rates on cross-modal and cross-spectral benchmark [10]. L2-Net† denotes results of L2-Net [23] with densely sampled windows.

	WTA optimization					SF optimization [13]				
	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	Average	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	Average
ANCC [47]	23.21	20.42	25.19	26.14	23.74	18.45	14.14	11.96	19.24	15.95
RSNCC [9]	27.51	25.12	18.21	27.91	24.69	13.41	15.87	9.15	18.21	14.16
SIFT [11]	24.11	18.72	19.42	27.18	22.36	18.51	11.06	14.87	20.78	16.31
DAISY [12]	27.61	26.30	20.72	27.41	25.51	20.42	10.84	12.71	22.91	16.72
BRIEF [34]	29.14	18.29	17.13	26.43	22.75	17.54	9.21	9.54	19.72	14.00
LSS [27]	27.82	19.18	18.21	26.14	22.84	16.14	11.88	9.11	18.51	13.91
LIOP [38]	24.42	16.42	14.22	20.42	18.87	15.32	11.42	10.22	17.12	13.52
DASC [10]	14.51	13.24	10.32	16.42	13.62	13.42	7.11	7.21	11.21	9.74
MatchN. [26]	19.72	16.54	20.81	27.14	21.05	17.51	10.82	11.84	12.34	13.13
DeepC. [68]	20.71	20.78	16.84	21.84	20.04	17.11	14.21	10.87	11.98	13.54
DeepD. [21]	16.72	17.81	12.72	20.71	16.99	14.87	10.88	12.87	13.93	13.14
Q-Net [40]	10.11	16.75	12.81	22.95	15.66	10.40	17.42	13.92	12.38	13.53
LIFT [22]	14.82	14.32	10.11	17.84	14.27	12.88	10.28	9.77	10.54	10.87
L2-Net [23]	13.79	13.16	9.92	19.11	13.99	11.92	15.22	11.20	11.69	12.51
L2-Net† [23]	12.61	14.22	10.22	20.54	14.40	10.51	14.66	10.90	12.17	12.06
FCSS [24]	11.87	**9.84**	**7.99**	17.64	11.84	12.10	6.28	6.11	10.84	8.83
SSC	**10.12**	10.12	8.22	**14.22**	**10.67**	**9.12**	**6.18**	**5.22**	**9.12**	**7.41**
DSC	**8.12**	**8.22**	**6.72**	**13.28**	**9.09**	**7.62**	**5.12**	**4.72**	**8.01**	**6.37**
GI-DSC	**9.30**	**7.92**	**6.86**	**12.92**	**9.25**	**7.12**	**4.75**	**4.42**	**7.06**	**5.84**

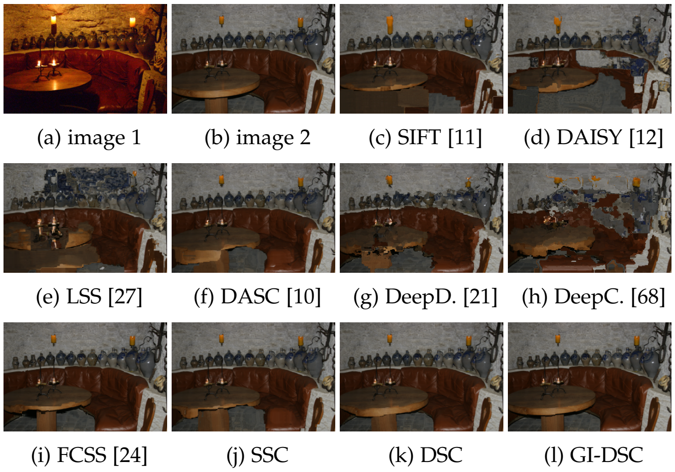

(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [27] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [68]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 13. Dense correspondence evaluations for flash-noflash image pairs on cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

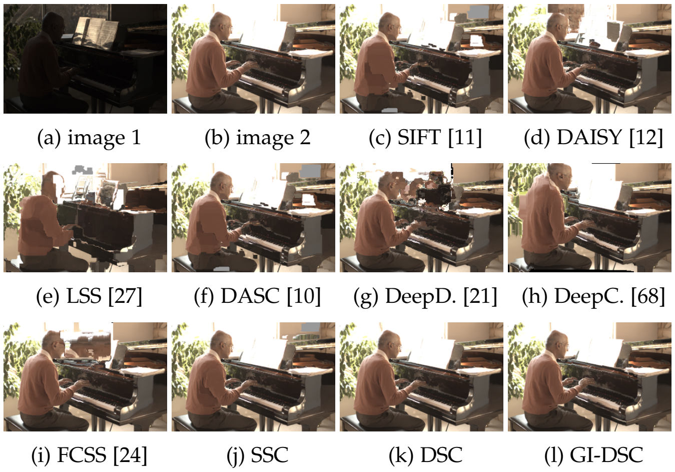

(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [27] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [68]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 14. Dense correspondence evaluations for different exposure image pairs on cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

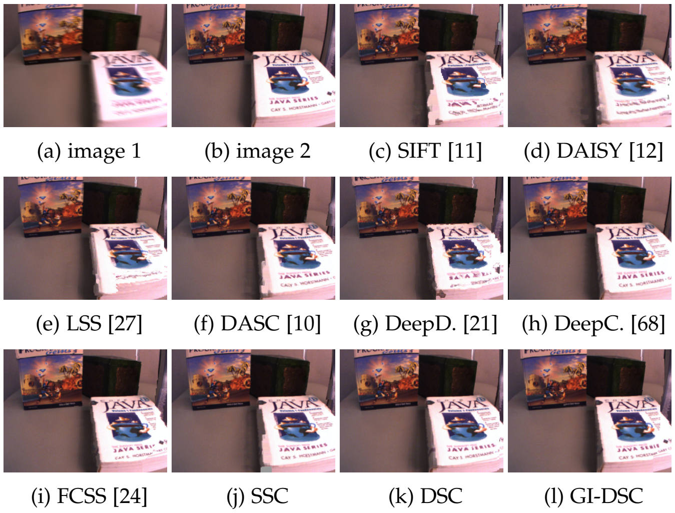
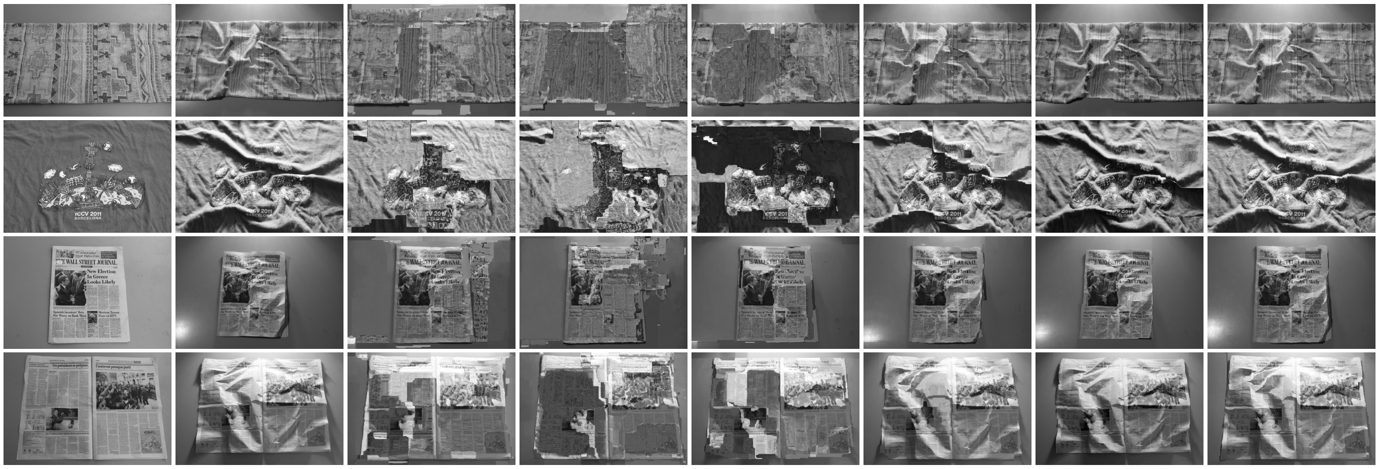

(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [27] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [68]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 15. Dense correspondence evaluations for blurred-shapre image pairs on cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

optimization. Note that since the geometric variation across stereo images exists only on translation field, GI-DSC was not evaluated in this experiment. Area-based approaches (ANCC [47] and RSNCC [9]) are sensitive to severe radiometric variations, especially when local variations occur frequently. Feature descriptor-based methods (SIFT [11], DAISY [12], BRIEF [34], LSS [27], and DASC [10]) perform better than the area-based approaches, but they also provide limited performance. Although the state-of-the-art CNN-based descriptor (MC-CNN [66]) has shown good results, it exhibits limited performance in cases of severe radiometric variation. Note that since other state-of-the-art stereo matching methods directly estimate disparity maps in an end-to-end manner, they were not evaluated in this experiment. Our DSC achieves the best results both quantitatively and qualitatively. Compared to SSC, the performance of DSC is highly improved, where the performance benefits of the hierarchical structure are apparent.

5.4 Cross-modal and Cross-spectral Benchmark

We evaluated DSC and GI-DSC on a cross-modal and cross-spectral benchmark [10] containing various kinds of image pairs, namely RGB-NIR, flash-noflash, different exposures,



(a) image 1 (b) image 2 (c) DAISY [12] (d) BRIEF [34] (e) LSS [27] (f) DaLI [31] (g) FCSS [24] (g) DSC

Fig. 16. Dense correspondence evaluations for images with different illumination conditions and non-rigid image deformations [31]. The source images were warped to the target images using correspondences.

TABLE 2
Average error rates on the DaLI benchmark [31].

Methods	deform.	illum.	deform./ illum.	Average
DAISY [12]	43.98	42.72	43.42	43.37
BRIEF [34]	41.51	37.14	41.35	40.00
LSS [27]	40.81	39.54	40.11	40.12
LIOP [38]	28.72	31.72	30.21	30.22
DaLI [31]	27.12	27.31	27.99	27.47
DASC [10]	26.21	24.83	27.51	26.18
VGG [16]	25.72	23.41	22.51	23.88
LIFT [22]	27.42	27.11	29.28	27.94
L2-Net [23]	26.34	25.74	26.84	26.31
FCSS [24]	**22.18**	24.72	**19.72**	**22.21**
SSC	23.42	**22.21**	24.17	23.27
DSC	**20.14**	**20.72**	**21.87**	**20.91**
GI-DSC	**18.47**	**16.25**	**18.24**	**17.65**

TABLE 3
Average error rates on the tri-modal human benchmark [32].

	RGB-depth		RGB-thermal		depth-thermal	
	LTA	IoU	LTA	IoU	LTA	IoU
DAISY [12]	45.51	0.41	36.31	0.44	53.21	0.52
BRIEF [34]	46.22	0.46	48.11	0.41	57.22	0.53
LSS [27]	49.27	0.52	49.38	0.42	51.87	0.42
LIOP [38]	41.75	0.37	48.27	0.36	50.78	0.39
DaLI [31]	40.99	0.39	48.72	0.43	53.95	0.50
DASC [10]	36.72	0.36	38.27	0.39	43.72	0.41
VGG [16]	33.16	0.39	38.11	0.42	46.72	0.38
LIFT [22]	38.72	0.47	43.51	0.49	50.84	0.53
L2-Net [23]	36.27	0.41	38.84	0.38	42.54	0.47
FCSS [24]	30.82	0.31	**29.71**	**0.30**	**39.78**	**0.34**
SSC	**30.11**	**0.29**	30.87	0.31	42.81	0.36
DSC	**26.19**	**0.24**	**29.38**	**0.27**	**36.22**	**0.27**
GI-DSC	**22.63**	**0.19**	**27.42**	**0.24**	**30.82**	**0.21**

and blurred-sharp. Sparse ground-truths for those images are used for error measurement as done in [10].

Fig. 12, Fig. 13, Fig. 14, and Fig. 15 provide a qualitative comparison of the DSC and GI-DSC descriptors to other state-of-the-art approaches for RGB-NIR, flash-noflash, different exposures, and blurred-sharp images, respectively. As already described in the literature [9], gradient-based approaches such as SIFT [11] and DAISY [12] have shown limited performance for RGB-NIR pairs where gradient reversals and inversions frequently appear. BRIEF [34] cannot deal with noisy regions and modality-based appearance differences since it is formulated on pixel differences only. Unlike these approaches, LSS [27] and DASC [10] consider local self-similarities, but LSS suffers from limited discriminative power. DASC also exhibits limited performance due to the sensitivity of patch-wise receptive field pooling. State-of-the-art CNN-based descriptors such as MatchN. [26], DeepC. [68], DeepD. [21], LIFT [22], L2-Net [23], and FCSS [24], pretrained on non cross-modal image pairs, cannot provide reliable correspondence estimation performance on cross-modal matching. Even though those methods have shown high robustness to photometric variations, they provide limited precision in localization. Moreover, large-scale training datasets are lacking for learning those descriptors. Q-Net [40] trained on the RGB-NIR dataset [1] has shown limited generalization ability to the appearance variations of various modalities such as flash-noflash, different exposures, and blurred-sharp. Compared to those methods, DSC displays better correspondence estimation. We also performed a quantitative evaluation with results listed in Table 1, which also clearly demonstrates the effectiveness of DSC. Note that the geometric variation across images provided from the cross-modal and cross-spectral benchmark [10] is not substantial, and thus it is relatively difficult to show the effectiveness of GI-DSC in terms of handling the geometry variation. Nevertheless, even in the benchmark, GI-DSC demonstrates an improved matching performance over DSC by considering geometry-invariant receptive fields.

5.5 DaLI Benchmark

We also evaluated the DSC and GI-DSC descriptors on a publicly available dataset featuring challenging non-rigid deformations and severe illumination changes [31]. Fig. 16 presents dense correspondence estimates for this benchmark [31]. A quantitative evaluation is given in Table 2 using ground-truth feature points sparsely extracted for each image, although DSC and GI-DSC are designed to estimate dense correspondences. As expected, conventional gradient-based and intensity comparison-based feature descriptors, including SIFT [11], DAISY [12], and BRIEF [34], are relatively less effective on such images. LSS [27] and DASC [10] exhibit relatively high performance for illumination changes, but perform less well on non-rigid geometric

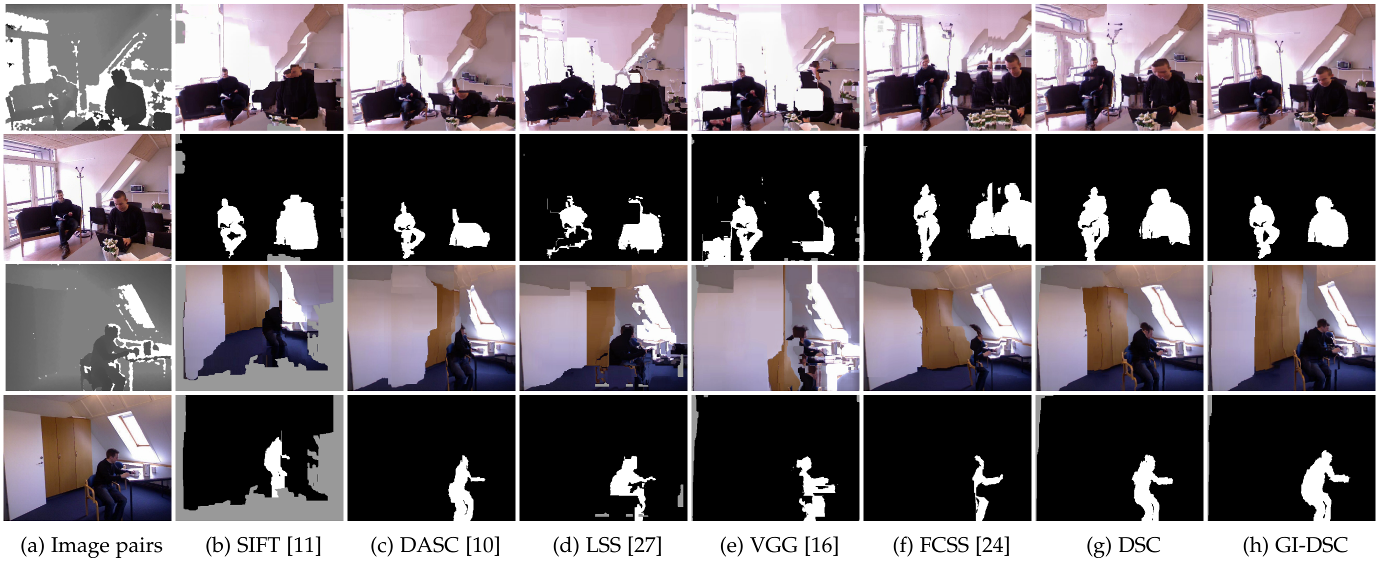


Fig. 17. Comparison of qualitative evaluation on RGB-depth human benchmark [32]. The results consist of warped color images and warped ground-truth human annotations.

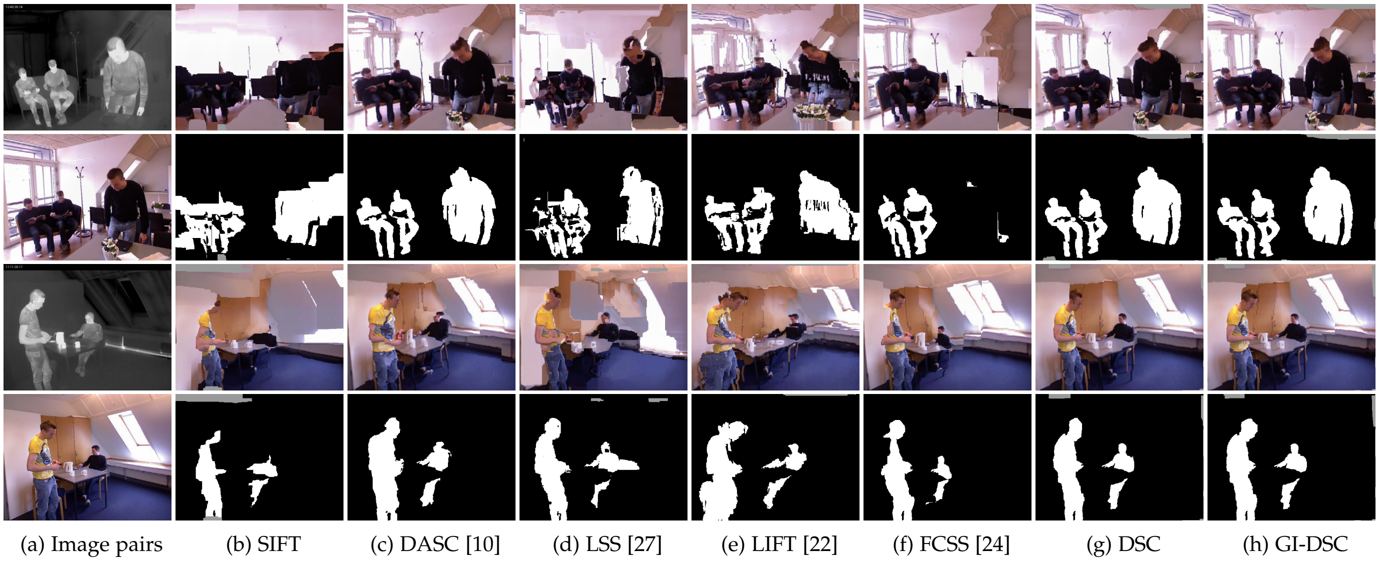


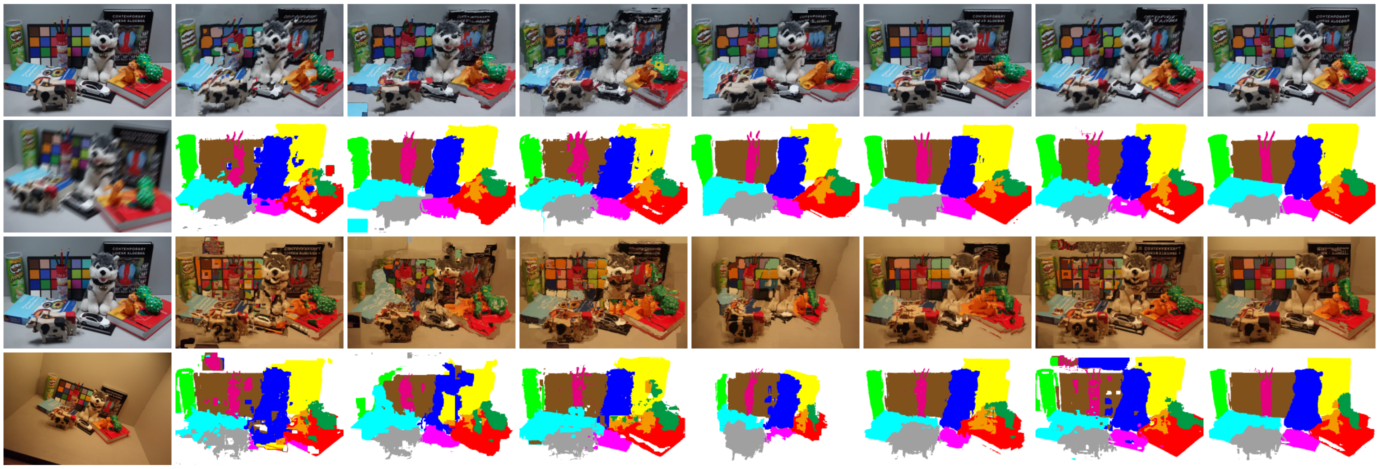
Fig. 18. Comparison of qualitative evaluation on RGB-thermal human benchmark [32]. The results consist of warped color images and warped ground-truth human annotations.

deformations. LIOP [38] provides robustness to radiometric variations, but is sensitive to non-rigid deformations. Although DaLI [31] provides robust correspondences, it requires considerable computation for dense matching. DSC offers greater discriminative power as well as more robustness to non-rigid deformations in comparison to the state-of-the-art cross-modality descriptors. State-of-the-art deep CNN-based methods such as FCSS [24] also show strong performance but require tremendous training time and a large number of training samples. Furthermore, by using explicit geometric estimation modules, GI-DSC present state-of-the-art performance for non-rigid deformations.

5.6 Tri-modal Human Benchmark

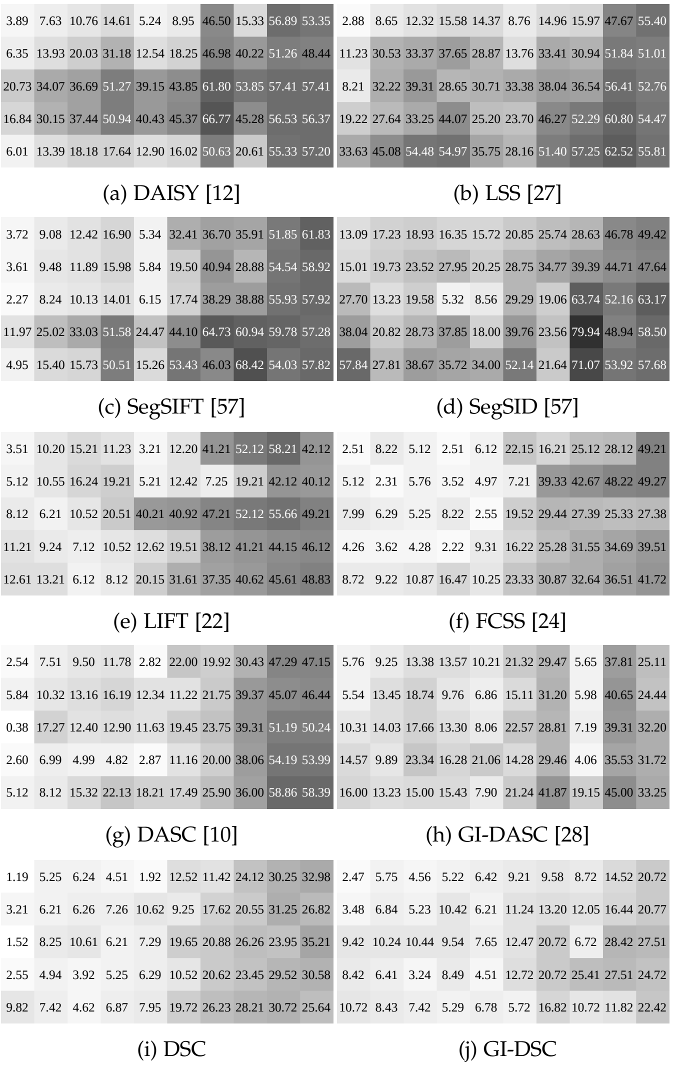
We additionally evaluated our descriptor on the tri-modal human body segmentation dataset [32] which includes RGB-Depth-FIR pairs. The dataset contains 11,537 frames divided into three indoor scenes and among them, 5,724 frames with human body annotations. To quantitatively measure the estimated correspondence quality, we use the label transfer accuracy (LTA) [13], [28] and intersection over union (IoU) metrics [14], [24] with manually built ground-truth object annotation maps, a practical alternative when no ground-truth correspondence is available. Note that different from LTA, IoU isolates the matching quality for foreground objects, separate from irrelevant background pixels.

Fig. 17 and Fig. 18 display qualitative comparisons for RGB-Depth and RGB-FIR pairs, respectively. Table 3 present a quantitative evaluation in terms of LTA and IoU. In comparison to the experiments of previous sections, this experiment uses RGB-Depth, RGB-FIR, and Depth-FIR pairs with more severe cross-modal variations. Similar to previous experiments, conventional handcrafted descriptors such as DASC [10] show limited performance. Although state-of-the-art CNN-based methods produce improvements, they cannot deal with non-rigid deformations or severe appearance variations across cross-modal images.



(a) Image pairs (b) DAISY (c) Seg-SIFT [57] (d) Seg-SID [57] (e) DASC [10] (f) GI-DASC [28] (g) DSC (h) GI-DSC

Fig. 20. Comparison of qualitative evaluation on DIML cross-modal benchmark [70]. The results consist of warped color images and warped ground-truth annotations.



3.89	7.63	10.76	14.61	5.24	8.95	46.50	15.33	56.89	53.35
6.35	13.93	20.03	31.18	12.54	18.25	46.98	40.22	51.26	48.44
20.73	34.07	36.69	51.27	39.15	43.85	61.80	53.85	57.41	57.41
16.84	30.15	37.44	50.94	40.43	45.37	66.77	45.28	56.53	56.37
6.01	13.39	18.18	17.64	12.90	16.02	50.63	20.61	55.33	57.20

(a) DAISY [12]

2.88	8.65	12.32	15.58	14.37	8.76	14.96	15.97	47.67	55.40
11.23	30.53	33.37	37.65	28.87	13.76	33.41	30.94	51.84	51.01
8.21	32.22	39.31	28.65	30.71	33.38	38.04	36.54	56.41	52.76
19.22	27.64	33.25	44.07	25.20	23.70	46.27	52.29	60.80	54.47
33.63	45.08	54.48	54.97	35.75	28.16	51.40	57.25	62.52	55.81

(b) LSS [27]

3.72	9.08	12.42	16.90	5.34	32.41	36.70	35.91	51.85	61.83
3.61	9.48	11.89	15.98	5.84	19.50	40.94	28.88	54.54	58.92
2.27	8.24	10.13	14.01	6.15	17.74	38.29	38.88	55.93	57.92
11.97	25.02	33.03	51.58	24.47	44.10	64.73	60.94	59.78	57.28
4.95	15.40	15.73	50.51	15.26	53.43	46.03	68.42	54.03	57.82

(c) SegSIFT [57]

13.09	17.23	18.93	16.35	15.72	20.85	25.74	28.63	46.78	49.42
15.01	19.73	23.52	27.95	20.25	28.75	34.77	39.39	44.71	47.64
27.70	13.23	19.58	5.32	8.56	29.29	19.06	63.74	52.16	63.17
38.04	20.82	28.73	37.85	18.00	39.76	23.56	79.94	48.94	58.50
57.84	27.81	38.67	35.72	34.00	52.14	21.64	71.07	53.92	57.68

(d) SegSID [57]

3.51	10.20	15.21	11.23	3.21	12.20	41.21	52.12	58.21	42.12
5.12	10.55	16.24	19.21	5.21	12.42	7.25	19.21	42.12	40.12
8.12	6.21	10.52	20.51	40.21	40.92	47.21	52.12	55.66	49.21
11.21	9.24	7.12	10.52	12.62	19.51	38.12	41.21	44.15	46.12
12.61	13.21	6.12	8.12	20.15	31.61	37.35	40.62	45.61	48.83

(e) LIFT [22]

2.51	8.22	5.12	2.51	6.12	22.15	16.21	25.12	28.12	49.21
5.12	2.31	5.76	3.52	4.97	7.21	39.33	42.67	48.22	49.27
7.99	6.29	5.25	8.22	2.55	19.52	29.44	27.39	25.33	27.38
4.26	3.62	4.28	2.22	9.31	16.22	25.28	31.55	34.69	39.51
8.72	9.22	10.87	16.47	10.25	23.33	30.87	32.64	36.51	41.72

(f) FCSS [24]

2.54	7.51	9.50	11.78	2.82	22.00	19.92	30.43	47.29	47.15
5.84	10.32	13.16	16.19	12.34	11.22	21.75	39.37	45.07	46.44
0.38	17.27	12.40	12.90	11.63	19.45	23.75	39.31	51.19	50.24
2.60	6.99	4.99	4.82	2.87	11.16	20.00	38.06	54.19	53.99
5.12	8.12	15.32	22.13	18.21	17.49	25.90	36.00	58.86	58.39

(g) DASC [10]

5.76	9.25	13.38	13.57	10.21	21.32	29.47	5.65	37.81	25.11
5.54	13.45	18.74	9.76	6.86	15.11	31.20	5.98	40.65	24.44
10.31	14.03	17.66	13.30	8.06	22.57	28.81	7.19	39.31	32.20
14.57	9.89	23.34	16.28	21.06	14.28	29.46	4.06	35.53	31.72
16.00	13.23	15.00	15.43	7.90	21.24	41.87	19.15	45.00	33.25

(h) GI-DASC [28]

1.19	5.25	6.24	4.51	1.92	12.52	11.42	24.12	30.25	32.98
3.21	6.21	6.26	7.26	10.62	9.25	17.62	20.55	31.25	26.82
1.52	8.25	10.61	6.21	7.29	19.65	20.88	26.26	23.95	35.21
2.55	4.94	3.92	5.25	6.29	10.52	20.62	23.45	29.52	30.58
9.82	7.42	4.62	6.87	7.95	19.72	26.23	28.21	30.72	25.64

(i) DSC

2.47	5.75	4.56	5.22	6.42	9.21	9.58	8.72	14.52	20.72
3.48	6.84	5.23	10.42	6.21	11.24	13.20	12.05	16.44	20.77
9.42	10.24	10.44	9.54	7.65	12.47	20.72	6.72	28.42	27.51
8.42	6.41	3.24	8.49	4.51	12.72	20.72	25.41	27.51	24.72
10.72	8.43	7.42	5.29	6.78	5.72	16.82	10.72	11.82	22.42

(j) GI-DSC

Fig. 19. Comparison of quantitative evaluation on DIML cross-modal benchmark [70]. Each result represents the LTA for geometric (x-axis) and photometric (y-axis) variations, respectively.

TABLE 4
Average error rates on the DIML cross-modal benchmark [70].

	photometry		geometry		all	
	LTA	IoU	LTA	IoU	LTA	IoU
DAISY [12]	36.42	0.42	48.42	0.51	42.42	0.47
BRIEF [34]	40.51	0.50	47.21	0.54	43.86	0.52
LSS [27]	38.51	0.42	47.80	0.43	43.16	0.43
LIOP [38]	26.71	0.36	52.03	0.41	37.37	0.39
DaLI [31]	34.71	0.34	49.82	0.39	42.27	0.37
DASC [10]	20.41	0.31	30.81	0.33	25.61	0.32
GI-DASC [24]	21.92	0.32	**21.84**	0.26	21.88	0.29
VGG [16]	22.07	0.29	41.11	0.27	31.59	0.28
LIFT [22]	23.11	0.30	42.02	0.31	32.57	0.31
L2-Net [23]	26.75	0.35	38.74	0.45	32.75	0.40
FCSS [24]	**18.72**	**0.27**	31.80	**0.24**	**25.26**	**0.26**
SSC	19.78	0.29	31.71	0.28	25.75	0.29
DSC	**16.72**	**0.24**	**26.11**	**0.25**	**21.42**	**0.25**
GI-DSC	**14.70**	**0.19**	**16.27**	**0.20**	**15.49**	**0.20**

5.7 DIML Cross-modal Benchmark

We further evaluated the DSC and GI-DSC descriptors on the DIML cross-modal benchmark [28] with both photometric and geometric variations. Ten geometry image sets were captured with geometric variations that arise from a combination of viewpoint, scale, and rotation differences, and each image set consists of images taken under five different photometric variation pairs including illumination, exposure, flash-noflash, blur, and noise. The DIML cross-modal benchmark thus consists of 100 images of size 1200 × 800. For quantitative evaluation, we used LTA and IoU, similar to Sec. 5.6. We follow the experimental configuration in [28], where for an image from the reference geometry image set, we estimate visual correspondence maps with images from other geometry image sets, and then compute LTA. Furthermore, visual correspondence maps are estimated for each photometric pair.

Fig. 19 exhibits the LTA for varying photometric and geometric deformations on the DIML cross-modal benchmark [28]. Table 4 and Fig. 20 presents quantitative and qualitative evaluation results, respectively. As expected, conventional gradient-based and intensity comparison-based feature descriptors, including SIFT [11], DAISY [12], and BRIEF [34], do not provide weaker correspondence performance. LSS [27] and DASC [10] exhibit relatively high performance for illumination changes, but are limited on non-rigid deformations. LIOP [38] provides robustness to radiometric variations, but is sensitive to non-rigid deformations. Although DaLI [31] provides robust correspondences,

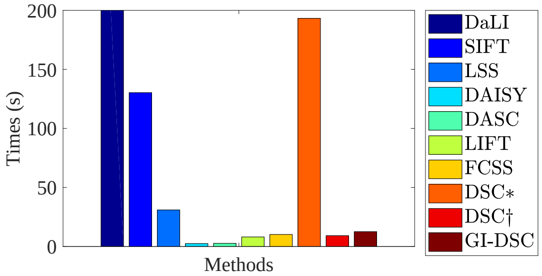


Fig. 21. Computation speed of DSC and GI-DSC descriptors and other state-of-the-art descriptors. The brute-force and efficient implementations of DSC are denoted by * and †, respectively.

it requires considerable computation for dense matching. State-of-the-art CNN-based descriptors such as LIFT [22] and FCSS [24] cannot deal with photometric and geometric variations simultaneously, resulting in limited performance. DSC offers greater discriminative power as well as more robustness to non-rigid deformation in comparison to the state-of-the-art cross-modality descriptors, but it remains vulnerable to severe geometric variations. Unlike these, GI-DSC shows robustness to both photometric and geometric variations.

5.8 Computational Speed

In Fig. 21, we compare the computation speed of DSC and GI-DSC to the state-of-the-art descriptors. Although deep CNN-based descriptors such as LIFT [22] and FCSS [24] are efficient at testing time compared to handcrafted descriptors such as DaLI [31], SIFT [11], and LSS [27], they entail a large computational burden at training time and require a large number of training samples. Compared to the brute-force implementation of DSC, the efficient implementation of DSC greatly reduces computation time. Moreover, compared to DSC, GI-DSC needs only marginal additional computation while providing high geometric invariance. Even though the DSC and GI-DSC descriptors need more computation compared to some previous dense descriptors, it provides significantly improved matching performance as described previously and is training-free.

6 Conclusion

The DSC and GI-DSC descriptors were proposed for establishing dense correspondences between images taken under different imaging modalities. Their high performance in comparison to state-of-the-art descriptors can be attributed to greater robustness to non-rigid deformations because of their effective pooling scheme, and more importantly their heightened discriminative power from a more comprehensive representation of self-similar structure and their formulation in a hierarchical manner. Over an extensive set of experiments that cover a broad range of cross-modal differences, DSC and GI-DSC were validated by their higher performance in comparison to existing handcrafted and deep CNN-based descriptors. Thanks to their robustness to non-rigid deformations and high discriminative power, DSC and GI-DSC can potentially be used to benefit object detection and semantic segmentation in future work.

References

[1] M. Brown and S. Susstrunk, "Multispectral sift for scene category recognition," *In: CVPR*, 2011.

[2] Q. Yan, X. Shen, L. Xu, and S. Zhuo, "Cross-field joint image restoration via scale map," *In: ICCV*, 2013.

[3] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," *In: CVPR*, 2015.

[4] D. Krishnan and R. Fergus, "Dark flash photography," *In: SIGGRAPH*, 2009.

[5] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes," *In: SIGGRAPH*, 2012.

[6] Y. HaCohen, E. Shechtman, and E. Lishchinski, "Deblurring by example using dense correspondence," *In: ICCV*, 2013.

[7] H. Lee and K. Lee, "Dense 3d reconstruction from severely blurred images using a single moving camera," *In: CVPR*, 2013.

[8] G. Petschnigg, M. Agrawals, and H. Hoppe, "Digital photography with flash and no-flash image pairs," *In: SIGGRAPH*, 2004.

[9] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," *In: ECCV*, 2014.

[10] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," *In: CVPR*, 2015.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[12] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. PAMI*, vol. 32, no. 5, pp. 815–830, 2010.

[13] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 815–830, 2011.

[14] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," *In: CVPR*, 2013.

[15] P. Pinggera, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," *In: BMVC*, 2012.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *In: ICLR*, 2015.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. PAMI*, vol. 36, no. 8, pp. 1573–1585, 2014.

[18] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional acitivation features," *In: ECCV*, 2014.

[19] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: A comparison to sift," *arXiv:1405.5769*, 2014.

[20] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *In: ICML*, 2014.

[21] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," *In: ICCV*, 2015.

[22] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," *In: ECCV*, 2016.

[23] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," *In: CVPR*, 2017.

[24] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," *In: CVPR*, 2017.

[25] J. Dong and S. Soatto, "Domain-size pooling in local descriptors: Dsp-sift," *In: CVPR*, 2015.

[26] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," *In: CVPR*, 2015.

[27] E. Schechtman and M. Irani, "Matching local self-similarities across images and videos," *In: CVPR*, 2007.

[28] S. Kim, D. Min, B. Ham, M. N. Do, and K. Sohn, "Dasc: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation," *IEEE Trans. PAMI*, vol. 39, no. 9, pp. 1712–1729, 2017.

[29] S. Kim, D. Min, B. Ham, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," *IEEE Trans. PAMI*, 2018.

[30] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, pp. 7–42, 2002.

[31] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, "Dali: Deformation and light invariant descriptor," *IJCV*, vol. 115, no. 2, pp. 136–154, 2015.

[32] C. Palmero, A. Clapes, C. Bahnsen, and A. Mogelmose, "Multimodal rgb-depth-thermal human body segmentation," *IJCV*, vol. 118, no. 2, pp. 217–239, 2016.

[33] S. Kim, D. Min, S. Lin, and K. Sohn, "Deep self-correlation descriptor for dense cross-modal correspondence," *In: ECCV*, 2016.

[34] M. Calonder, "Brief : Computing a local binary descriptor very fast," *IEEE Trans. PAMI*, vol. 34, no. 7, pp. 1281–1298, 2011.

[35] T. Trzcinski, M. Christoudias, and V. Lepetit, "Learning image descriptor with boosting," *IEEE Trans. PAMI*, vol. 37, no. 3, pp. 597–610, 2015.

[36] K. Alex, S. Ilya, and E. H. Geoffrey, "Imagenet classification with deep convolutional neural networks," *In: NIPS*, 2012.

[37] S. Saleem and R. Sablatnig, "A robust sift descriptor for multispectral images," *IEEE SPL*, vol. 21, no. 4, pp. 400–403, 2014.

[38] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," *In: ICCV*, 2011.

[39] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," *In: CVPR Workshop*, 2016.

[40] C. A. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," *In: Sensors*, vol. 17, no. 4, 2017.

[41] P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, V. Gleeson, S. Brady, and A. Schnabel, "Mind: Modality indepdent neighbourhood descriptor for multi-modal deformable registration," *MIA*, vol. 16, no. 3, pp. 1423–1435, 2012.

[42] A. Torabi and G. Bilodeau, "Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos," *PR*, vol. 46, no. 2, pp. 578–589, 2013.

[43] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *JPRS*, vol. 90, no. 7, pp. 83–95, 2014.

[44] J. Pluim, J. Maintz, and M. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Trans. MI*, vol. 22, no. 8, pp. 986–1004, 2003.

[45] Y. Heo, K. Lee, and S. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Trans. PAMI*, vol. 35, no. 5, pp. 1094–1106, 2013.

[46] J. Xu, Q. Yang, J. Tang, and Z. Feng, "Linear time illumination invariant stereo matching," *IJCV*, 2016.

[47] Y. Heo, K. Lee, and S. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. PAMI*, vol. 33, no. 4, pp. 807–822, 2011.

[48] M. Irani and P. Anandan, "Robust multi-sensor image alignment," *In: ICCV*, 1998.

[49] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," *In: ICCV*, 2013.

[50] T. Hassner, V. Mayzels, and L. Zelnik-Manor, "On sifts and their scales," *In: CVPR*, 2012.

[51] W. Qiu, X. Wang, X. Bai, A. Yuille, and Z. Tu, "Scale-space sift flow," *In: WACV*, 2014.

[52] J. Hur, H. Lim, C. Park, and S. C. Ahn, "Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation," *In: CVPR*, 2015.

[53] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," *In: ECCV*, 2010.

[54] H. Yang, W. Lin, and J. Lu, "Daisy filter flow: A generalized discrete approach to dense correspondences," *In: CVPR*, 2014.

[55] J. Lu, H. Yang, D. Min, and M. N. Do, "Patchmatch filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation," *In: CVPR*, 2013.

[56] I. Kokkinos and A. Yuille, "Scale invariance without scale selection," *In: CVPR*, 2008.

[57] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. M. Noguer, "Dense segmentation-aware descriptors," *In: CVPR*, 2013.

[58] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger, "Robust anisotropic diffusion," *IEEE Trans. IP*, vol. 7, no. 3, pp. 421–432, 1998.

[59] E. Gastal and M. Oliveira, "Domain transform for edge-aware image and video processing," *In: SIGGRAPH*, 2011.

[60] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. PAMI*, vol. 35, no. 6, pp. 1397–1409, 2013.

[61] L. Seidenari, G. Serra, A. D. Bagdanov, and A. D. Bimbo, "Local pyramidal descriptors for image recognition," *IEEE Trans. PAMI*, vol. 36, no. 5, pp. 1033–1040, 2014.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. PAMI*, vol. 37, no. 9, pp. 1904–1916, 2015.

[63] K. Chatfield, J. Philbin, and A. Zisserman, "Efficient retrieval of deformable shape classes using local self-similarities," *In: ICCV Workshop*, 2009.

[64] M. Tau and T. Hassner, "Dense correspondences across scenes and scales," *IEEE Trans. PAMI*, vol. 38, no. 5, pp. 875–888, 2016.

[65] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. IP*, vol. 23, no. 12, pp. 5638–5653, 2014.

[66] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.

[67] H. O. Song, Y. Xiang, S. Jegelk, and S. Savarese, "Deep metric learing via lifted structured feature embedding," *In: CVPR*, 2016.

[68] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *In: CVPR*, 2015.

[69] Y. Boykov, O. Yeksler, and R. Zabih, "Fast approximation enermgy minimization via graph cuts," *IEEE Trans. PAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.

[70] online., http://diml.yonsei.ac.kr/∼srkim/DASC/.

Seungryong Kim received the B.S. and Ph.D. degrees in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2012 and 2018, respectively. He is currently a Post-Doctoral Researcher in Electrical and Electronic Engineering at Yonsei University. His current research interests include 2D/3D computer vision, computational photography, and machine learning, in particular, sparse/dense feature descriptor and continuous/discrete optimization.

Dongbo Min received the B.S., M.S., and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea, in 2003, 2005, and 2009, respectively. He was with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, as a Post-Doctoral Researcher from 2009 to 2010. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore. Since 2015, he has been an Assistant Professor with the Department of Computer Science and Engineering, Chungnam National University, Daejeon, Korea. His current research interests include computer vision, 2D/3D video processing, computational photography, augmented reality, and continuous/discrete optimization.

Stephen Lin received the B.S.E. degree in electrical engineering from Princeton University, NJ, and the Ph.D. degree in computer science and engineering from the University of Michigan, Ann Arbor. He is a Senior Researcher with the Internet Graphics Group, Microsoft Research Asia. His research interests include computer vision, image processing, and computer graphics. He served as a Program Co-Chair of the International Conference on Computer Vision 2011 and the Pacific-Rim Symposium on Image and Video Technology 2009.

Kwanghoon Sohn received the B.E. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1983, the M.S.E.E. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992. He was a Senior Member of the Research engineer with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, Korea, from 1992 to 1993, and a Post-Doctoral Fellow with the MRI Center, Medical School of Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently an Underwood Distinguished Professor with the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing and computer vision.